Open ASR for Icelandic: Resources and a Baseline System

Anna Björk Nikulásdóttir, Inga Rún Helgadóttir, Matthías Pétursson, Jón Guðnason
Center for Analysis and Design of Intelligent Agents, Language and Voice Lab (https://lvl.ru.is/)
Reykjavik University, Iceland

REYKJAVIK UNIVERSITY

Overview

Description of available language resources and their processing for use in a large vocabulary speech recognition (LVSR) system for Icelandic:

Málrómur speech corpus
Icelandic text corpus from the *Leipzig Corpora Collection*
Icelandic pronunciation dictionary

Experiments with different sizes and compositions of the acoustic training corpus show the need for a larger speech corpus and indicate that an ASR system could profit from carefully selected phonotactical data.

The ASR system and its sources and resources are open and freely available.[1,2,3]

Málrómur Speech Corpus

Recordings of a collection of sentences and isolated words from over 550 speakers of both genders and of various age.

Contains about 122,000 utterances or approx. 150 hours of speech.

The list of approx. 33,000 prompts contains several categories: news articles, proper names, Icelandic street names and locations, international cities and countries, numbers, URLs, miscellaneous words, and a list of carefully selected sentences regarding diphone and triphone distribution *(phonotactical sentence list*).

The distribution of the prompt categories is different between the list of prompts and the speech corpus: entries are recorded multiple times, leading to substantially larger partitions of URLs and phonotactical sentences in the speech corpus than in the prompts list (Fig. 1).

The repetition of prompts in general is not evenly distributed in the speech corpus, a substantial number of prompts is only recorded once while a few prompts are recorded more than 100 times.

Repetitions in speech corpus	No. of prompts	No. of recordings
1x	7,450	7,450
2x	8,751	17,502
3-4x	11,090	37,428
5-10x	4,671	28,822
11-20x	937	12,866
21-50x	82	2,321
51-100x	68	6,001
>100x	86	9,495
TOTAL	33,135	121,822

Prompt categories

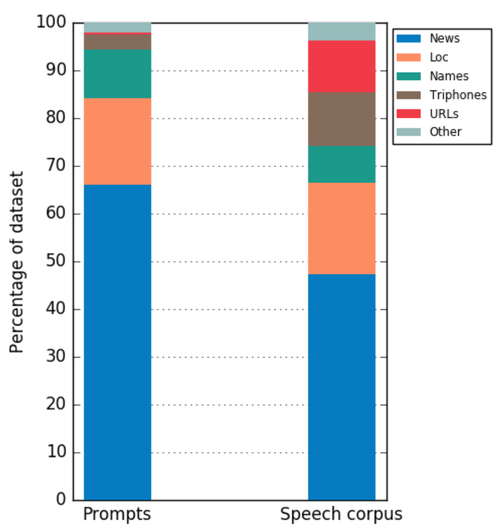


Figure 1: Distribution of prompt categories in the prompts list (~33k entries) and the Málrómur speech corpus (~122k entries).

Prompt length

The prompts from news articles and the phonotactical sentences are multi word sentences but most of the prompts from the other categories are single word utterances.

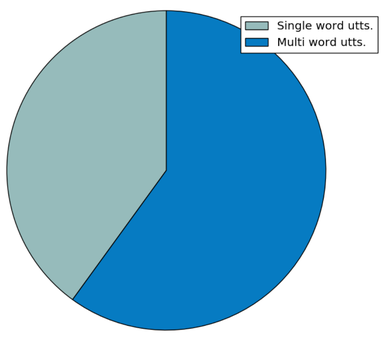


Figure 2: 40% of the recordings in Málrómur are single word utterances and 60% are multi word utterances. The average length of a multi word utterance is 5.3 words.

Phonotactically selected prompts

Improve diphone and triphone coverage in the speech corpus.

Increase effort in the creation of the prompts list.

Can be difficult to read due to rare words and unusual word combinations.

Can diminish the quality of the speech corpus measured in match between spoken utterances and their corresponding prompts.

Evaluation of Málrómur

Four evaluators rated 3,000 random recordings from Málrómur using ratings from 1 (=bad) to 4 (=very good). The kappa (Fleiss) value was 0.598 for the evaluation results, 72% of the recordings were rated as 'very good' by all evaluators.

Speakers identified whose recordings generally match the corresponding prompts.

Utterances from a random selection of the highly rated speakers used for test and development sets, leaving out utterances from the collection of phonotactical sentences and URLs.

LCC Text Corpus

An Icelandic text corpus, consisting mainly of news text, from the Leipzig Corpora Collection.

Sentence collection of 180 million tokens used for language modeling.

Basic text normalization: non-spoken symbols removed, common acronyms and abbreviations replaced.

Pronunciation Dictionary

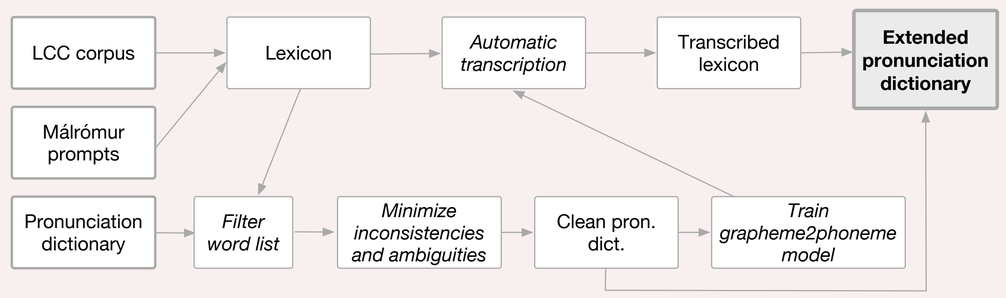
An initial version of an Icelandic pronunciation dictionary, created in 2003.

Compiled from news, novels and other sources.

65,000 entries transcribed by hand in IPA and SAMPA transcriptions.

Out-of-vocabulary rate of Málrómur is 19.5% with only 54% of the types occurring in the dictionary.

Lexicon Processing



Extended Lexicon

Fixed set of 58 IPA symbols for transcriptions: inconsistencies in manual transcriptions corrected.

Current version contains 136,000 entries, resulting in reduced OOV-rate of the Málrómur corpus: 8.9%.

76.2% of all types in Málrómur now included in the dictionary.

Baseline ASR System

Kaldi ASR toolkit used for implementation.

TDNN-LSTM network with ten hidden layers.

Trigram language models with modified Kneser-Ney smoothing.

Two test sets: open vocabulary and closed vocabulary (a subset of the open vocabulary set).

OOV rate of the open vocabulary test set: 5.7%.

Results

Training data	Open vocabulary	Closed vocabulary
Full training set (109h)	15.72% WER	7.99% WER
Subset including phonotactical sentences (95h)	16.17% WER	8.41% WER
Subset not including phonotactical sentences (93h)	16.27% WER	8.60% WER

Conclusion

The coverage of the lexicon has to be further improved for an LVSR system for a highly inflective language like Icelandic.

More acoustic training data is likely to reduce WER.

The effort of collecting and recording phonotactically selected texts should be balanced against the effort of recording more randomly collected data when constructing a speech corpus for low resourced languages.

Resources

[1]https://github.com/cadia-lvl/ice-asr/tree/master/ice-kaldi
[2]http://malfong.is
[3]https://tal.ru.is/

Center for Analysis and Design of Intelligent Agents
Gervigreindarsetur Háskólans í Reykjavík
rannís